JOURNAL OF THE EXPERIMENTAL ANALYSIS OF BEHAVIOR 1978, 29, 17-25 NUMBER 1 (JANUARY)

STABILITY CRITERIA

PETER R. KILLEEN[1]

ARIZONA STATE UNIVERSITY

Three approaches to the determination of behavioral stability were examined. In the first, a learning curve was fit to acquisition data (from Cumming and Schoenfeld, 1960), and the "experiment" stopped when the data approached sufficiently close to the theoretical asymptote. In the second, the data were analyzed for variability and linear and quadratic trend. In the third, the experiment was stopped when the magnitude of the daily changes in the data fell below a criterion. Accuracy was measured as deviation between the average value of the dependent variable when the experiment was stopped, and the average value over the last 100 sessions. The first approach was most accurate, but at the cost of requiring the most sessions and being the most difficult to apply. Both the second and third approaches provided acceptable criteria with a reasonable cost-accuracy tradeoff. The second approach permits a continuous adjustment of the criteria to accommodate the variability intrinsic in the experimental paradigm. The third, nomothetic, approach also takes into account the decreasing marginal utility of extended training sessions.

Key words: learning curves, trend analysis, nomothetic criteria, optional stopping

Cumming and Schoenfeld's words introduce this research as well as they did their own:

> In the literature, the term "stability" appears to refer to one or both of two things. In some places, it means that behavior is no longer changing significantly because it is close to its asymptotic value under the given conditions. In other contexts, the term seemingly refers to behavior . . . that shows minimal variability. . . The concern of the worker is carried by several concrete questions. . . : When could the experiment have been stopped with any desired probability that no further change in the dependent variable would have been observed? . . .What is a satisfactory rationale for defining "stability," and what is a reasonable criterion to set for accepting behavior as "stable"? (1960, p. 71).

The issue is important, for it is the custom of behavior analysts to continue an experiment until the data "appear stable", rather than terminate after a fixed number of sessions. Yet, to this date, there are no widely accepted criteria for stopping. The criterion tested by Cumming and Schoenfeld (1960) did not work:

> A 6-day period was considered to have met the stability criterion if the difference between the mean rate for the first three days and the rate for the second three days was no greater than 5% of the overall 6-day mean. . . Although the criterion itself tended on repeated application to select 6-day means at random, the use of the first occasion on which the criterion is met proves a bad choice in practice (pp. 78, 79).

Lacking algorithms for stopping experimental sessions, experimenters have usually relied on visual inspection to determine stability. Visual stability tests are presumably one of the contingency-shaped behaviors that are acquired during graduate education. Because such learning is often several stages removed from the relevant contingencies, we seek here to replace it with rule-governed behavior.

[1]This work was supported in part by grant BMS 74-23566 from the National Science Foundation. Listings of Basic computer programs relevant to this article may be obtained from the author. I am indebted to the Glass Bead Group for their cooperation in Experiment II and to Stephen Hanson for comments on the manuscript. Listings and reprints may be obtained from Peter R. Killeen, Department of Psychology, Arizona State University, Tempe, Arizona 85281.

In their experiment, Cumming and Schoenfeld (1960) trained six pigeons on a t-τ schedule (analogous to an FI 28.5-sec LH 1.5-sec schedule) for 200 sessions. These data, similar in appearance to other learning curves published in this journal, provide a good testing ground for alternate types of stopping criteria. The present paper evaluates three strategies for generating such criteria.

EXPERIMENT I

"Asymptote" is a line that a tangent to a curve approaches, as the curve is extended to infinity. To presume that behavior may be at some point asymptotic is to presume that the "learning curve" has a single asymptote, and that it will approach reasonably close to it within the lifetime of the subjects. Experimental contingencies may, however, establish an undamped negative feedback loop, so that the only stability is a stable oscillation between two asymptotes ("metastability"). Barring such a complication—or given indices such as frequency of oscillation rather than raw data—we may ask whether any simple quantitative model of the learning process provides useful estimates of asymptote and rate of approach to it.

There are a number of different models of learning curves to choose from: autocatalytic (Robertson, 1920), cumulative normal (Culler and Girden, 1951), power functions (Stevens and Savin, 1962), and exponential-integral (Anderson, 1963; Estes, 1950). There should be few important practical differences among the models' estimates of proximity to asmyptote, so we choose one of the simplest, the exponential learning curve:

$$R = A(1 - e^{-J/C}) \qquad (1),$$

where R is the dependent variable, A is the asymptote, e the base of the natural logarithms, J the number of sessions, and C a rate constant. The parenthetical expression ranges between 0 (at Session 0) and 1.0 (as J approaches infinity). The rate of approach to asymptote is governed by C, called the time-constant of the system. When $J = C$, the system is 63% of the way to asymptote; at $J = 3C$, 95%, and at $J = 5C$, 99%. Here is a simple model that provides direct measure of asymptote and rate of approach. But two parameters are seldom adequate, since the baseline from which change is to be measured is often not zero. This may be easily taken into account by adding a parameter "B" to the right side of the equation, which sets the dependent variable at level B (*i.e.,* baseline level) on the zeroth session.

Method

Subjects

The data were response rates collected by Cumming and Schoenfeld (1960) and are available from American Documentation Institute as Document No. 6244.

Apparatus

A PDP-11 computer was used.

Procedure

After each session, a simple iterative program searched for the best values of A and C. "Best" was taken as that value which minimized the sum of squared deviations between the obtained data and the theoretical learning curve. B (response rate of Session "0") was fiixed at a value extrapolated from the first two sessions. A number of stopping criteria were evaluated, with their merit decided by the smallness of the deviation between the average rate during the six days before stopping and the average rate over Sessions 80 to 180 (the last stable 100 sessions, labelled "Asymptote" in Table 1).

Results

The data were considered stable when two conditions were satisfied: the learning process had to be 99% of the way to completion (*i.e.,* $J \geq 5C$) and the average of the last six sessions had to be within 5% of the predicted asymptote. Table 1 shows the session in which the criterion was reached by each of the animals, the average rate for the six days preceding that session, and the per cent by which this average deviates from that of the last 100 sessions.

Discussion

The results hardly justify the effort put into curve-fitting and updating that fit every session. The average error was 14% (the average error between the theoretical asymptotes and the last 100 sessions was also 14%). While this is considerably better than the performance of Cumming and Schoenfeld's criterion (average

deviation: 25%), the average number of sessions required was also considerably greater. If we had indiscriminately stopped all animals on the forty-third session, the average individual deviation would have been 12%. Systematic evaluation of other criteria ($J \geq 2C$ through $J \geq 6C$, deviation from theoretical asymptote from 2% through 10%, χ^2, *etc.*) yielded none that fared better.

The reason for the failure of this technique is the failure of the pigeons to approach an asymptote in a smooth, monotonic fashion. A fast start, as was the case with the first two pigeons, would cause an overestimation of asymptote during the early sessions. This bias would be eliminated as the data continued to accumulate, but it was most likely that the criterion would be satisfied by data that were anomalously high and met the descending theoretical asymptote before it had stabilized at a more representative level.

Control theory might provide a better model for the stabilization of data than does learning theory, for it provides explicit representation for data that overshoot their asymptotic level. But that requires three additional parameters: the damping ratio, and the frequency and phase of oscillation. These make computer convergence difficult, and complicate the model to the point of impracticality. Alternately, if data from the earliest part of the experiment are expendable, we may improve our estimate of asymptote at the cost of an inferior account of the first few sessions. But the issue of how much data to discard itself stands in need of a criterion. That decision leads into the issue of exponential weights for the data, which, in this context, becomes as complicated as the control model. Other learning curves (*e.g.*, logistic) with other criteria for the elimination of data and the judgement of stability might yet prove viable, but we now proceed to other types of criteria.

EXPERIMENT II

What do people look at when they "eyeball" data for stability? In the following experiment, we attempted to distill an algorithm from a number of judgements of "stable by visual inspection".

Subjects

Five sophisticated laboratory researchers served.

Apparatus

Fifty-five graphs of "data points" were generated by computer from random variables with a mean of 50 and a standard deviation of five (scale: 25 units to the inch). Each graph consisted of six unconnected points, with no numerical ordinates visible.

Procedure

Each graph was displayed through a window in a file folder. Subjects were told to suppose that these were the last six sessions from an experiment that had been in progress for 25 sessions. Only one of the subjects knew the true nature of the data. They were asked whether they would stop the experiment at this point, and asked to rate their confidence in that decision on a scale of one to three.

Table 1

A comparison of various stability criteria. "Dev" is absolute per cent deviation between rates and asymptotes. The two values for "Avg Dev" are the average of the column and the deviation of the average rates from the average asymptotes.

		Criterion								
		Learning Curve			*Trend Analysis*			*Nomothetic Criterion*		
Subject	*Asymptote*	*Day*	*Rate*	*Dev*	*Day*	*Rate*	*Dev*	*Day*	*Rate*	*Dev*
26	30.7	60	40.7	33	20	45.2	47	22	43.0	40
27	76.0	56	83.8	10	14	54.4	28	18	59.7	21
28	36.7	36	40.3	10	29	38.3	4	32	36.0	2
33	42.3	57	38.5	9	13	38.3	9	16	44.8	6
42	46.1	26	54.5	18	8	58.0	26	14	57.0	24
45	44.6	21	45.7	2	43	48.0	8	35	51.8	16
Avg	46.1	43	50.6	14/9.7	21	47.0	20/2.0	23	48.7	18/5.7

Two subjects looked at the graphs in one order, three in the opposite order. The first five ratings were not included in the analysis.

Results and Discussion

Although the population from which these random variables were sampled had a 10% coefficient of variation and zero linear and quadratic trend, random sampling of only six items from the population generated a range of data configurations that resembled typical data, with reasonable disparity in both dispersion and trend from one graph to another.

Most subjects lamented their inability to see data from sessions previous to the six on display. In this sense, the experiment did not permit the full range of observing behavior typically involved in visual estimates of stability. The subjects reported that they attended to both variability and trend in making their decisions. A number of indices of variability and trend were therefore examined in an attempt to capture the essence of the subject's behavior. The most successful indices were the coefficient of variation, the amount of linear trend and the amount of quadratic trend, with both of the latter measured by weighting the data with appropriate orthogonal polynomials.

The coefficient of variation was calculated by dividing the sample standard deviation by the mean of the data, and the indices of trend were calculated by multiplying each of the scores by the appropriate weight (−5, −3, −1, 1, 3, 5, for linear trend, 5, −1, −4, −4, −1, 5 for quadratic trend), and dividing the weighted sum by the root mean square average of the weights (8.37 for linear, 9.17 for quadratic), and by the mean of scores. The correlations between the average confidence rating for each graph and each of the indices were: coefficient of variation, 0.73; linear trend, 0.58; quadratic trend, 0.53. The multiple correlation was 0.81.

Despite its success, the multiple regression may not be a good analog of the behavior of the raters. Rather than estimating a weighted sum of the indices, the subjects seemed to employ a noncompensatory approach: too much variability, or too much linear or quadratic trend would preclude stopping, no matter how close to zero the other indices were. We can reconstruct the binary decisions from the data with the following *post-hoc* analysis: continue running whenever the coefficient of variation exceeds 0.14, the coefficient of linear trend 0.12, or the coefficient of quadratic trend 0.20. These cut-points would have stopped the experiment every time the average rater stopped it, and continued the experiment 93% of the time the rater continued it.

Next, I applied the criterion to Cumming and Schoenfeld's data, stopping a subject whenever all three coefficients over a six-day period were less than 14%. Given that criterion, on the average, 21 sessions are required for stability, with an average deviation between rates over the last six sessions and asymptotic rates of 20% (see Table 1). This error falls between the 25% average deviation of Cumming and Schoenfeld's criterion, and the 14% average deviation of the "learning curve" criterion. Thus, while the "trend analysis" is better than Cumming and Schoenfeld's criterion, it leaves room for improvement. Some of the error arises from shifts in response rates late in training, and it would be difficult for any criterion imposed earlier in training to get past periods of relative stability that preceded the later changes in level. Another source of inaccuracy is sampling error: the average individual standard deviation during the last 100 sessions was 6.6 responses per minute. The standard error of the mean for samples of six sessions was therefore 2.7 responses per minute. It follows that the average deviation of sample means from population means (*i.e.*, rates over the last 100 sessions) will be about 4%. Response rates during Sessions 75 to 80, which Cumming and Schoenfeld considered asymptotic, deviated by an average of 9%. Four to 9% is therefore about the best we can hope for from these data.

Is it possible to approach closer to the floor of 4 to 9% with trend analyses? Tests employing larger numbers of sessions (10) and a range of different cutpoints failed to improve substantially on the 20% error. As Sidman (1960) noted, increasing the stringency of our criteria will not guarantee an increase in the stability of the data when those criteria are met; it may just cause the researchers "to spend a lifetime, if they are that stubborn, on the same uncompleted experiment. . . . Even if the criterion were occasionally met by chance, in the course of uncontrolled variability, the data would be chaotic. As a result, either the experiment will be abandoned (with

an attendant loss of time and effort) or the data will be invalid" (p. 260).

Sidman's words call our attention to two factors involved in every reasonable decision concerning stability: the levels of variability generally expected within a paradigm and within a laboratory, and the cost-effectiveness of additional training sessions. Whereas the cut-points of the trend analyses and the number of sessions it encompasses may be easily and properly adjusted by each user to allow for intrinsic variability,[2] it is only with the nomothetic approach that the costs of experimentation are explicitly taken into account.

EXPERIMENT III

"*N* should refer to the number of observations, not to the number of subjects", I was recently reminded. This sentiment, uttered in defense of "small *N*" research, implies a possible trade-off between the number of subjects and the number of sessions involved in an experiment. Different types of information are derived by conducting numerous sessions involving a few subjects, *versus* conducting a few sessions involving numerous subjects. In the former case, we achieve a good specification of the behavior of a few organisms, but little information about how representative they are of their species. In the latter case, we achieve a good sample of the population, but the behavior measured may be far from asymptotic.

It is our thesis that these *two types of information are both necessary,* and furthermore, that *they are commensurable.* We assume a nomothetic philosophy, according to which data from individual organisms are to be evaluated in terms of their contribution to our knowledge of population characteristics (Falk, 1956). Both sources of errors—deviation of individuals from their asymptotic performance, and deviation of that asymptotic performance from the performance of the typical animal of the species—are to be minimized.

The nomothetic assumption is invoked, not only because of its inherent reasonableness, but because it provides the needed standard for our next stability criterion. We know that as the sample size (*N*) is increased, the expected deviation between the sample mean and the population mean will decrease as the square root of N − 1. Similarly, as the number of sessions is increased, the deviation between the subjects' measured performance and their asymptotic performance will decrease. Given fixed resources, we can maximize the precision of our estimate of the population mean by judiciously allocating those resources between sessions and subjects. Conversely, given a decision about the number of subjects to be employed, we can stipulate when the decreasing marginal utility of additional sessions passes a threshold beyond which that decision should be revised—or, standing by that decision, when it becomes reasonable to terminate the experiment. This strategy does not assume that we have any particular interest in between-group comparisons; it does assume that we are interested in generalizing the results obtained with a sample to the population as a whole.

Method

To effect the analysis, we must estimate the probable magnitude of error that might arise both from sampling error and from failure of the subjects to reach asymptote. In the data reported by Cumming and Schoenfeld, the mean of response rates for the six subjects over the last 100 sessions was 46.1, with an estimated σ of 15.7. Assuming a normal distribution of error, we may infer that approximately 95% of the replications of their experiment with six subjects will generate asymptotic mean response rates within two standard errors of the mean, that is, between 33 and 59 responses per minute. The more subjects that are run, the more tightly these limits may be drawn. The standard error of the mean, calculated by dividing the sample estimate of the population standard deviation (s) by the square root of N, thus provides the needed relationship between sample size and expected sampling error.

Let us estimate the change in error that occurs with each additional session (dE/dJ) by subtracting the response rate during one session from that measured during the previous sessions. A provisional criterion for stability is the following: when the overall decrease in error derived from additional sessions becomes less than the decrease in error derivable from additional subjects (dE/d*N*), it becomes rea-

[2]We now routinely employ eight-session trend tests with cut-points around 15%.

sonable to terminate those subjects and initiate new subjects in their place. Alternatively, taking the original number of subjects as an index of the tolerable error—and this will be the usual course of action—it becomes reasonable to terminate the experiment. This formulation may be improved by selective application of the criterion: whenever the daily decrease in error for an individual subject becomes less than the criterion, terminate it and allocate the saved resources either to new subjects, or to more sessions for the slower subjects.

The above statement ignores the cost involved in conducting additional sessions or running additional subjects. If we take the unit cost to be one subject-session, the cost of each additional session is 1 (or N for the group). The cost of each additional subject must take into account the likelihood that it will be necessary to run the subject as long as the one it is replacing to obtain adequate stability—let us say J sessions. Our criterion then becomes: terminate a subject whenever

$$\frac{-dE}{dJ} \Big/ J \leq \frac{-dE}{dN} \Big/ 1. \tag{2a}$$

That is, terminate whenever the error reduction expected from one additional session, divided by the number of sessions to date, becomes less than the error reduction expected from one additional subject. This explicit introduction of the cost for additional sessions saves us from the Sisyphean fate of Sidman's stubborn researcher, for the accumulating costs of additional sessions progressively relaxes the criterion. The *ad-hoc* decision "long enough" (*i.e.*, visual stability) is replaced by a continuous and specifiable adjustment of the criterion.

If the structure or economics of the laboratory dictate termination of the group as a whole, Equation 2a becomes

$$\frac{-dE}{dJ} \Big/ J \leq \frac{-dE}{dN} \Big/ N, \tag{2b}$$

where the group is treated as a single organism, and the dE/dJ is based on the change in the dependent variable averaged over the group. The derivative of the standard error of the mean with respect to N is estimated by

$$\frac{dE}{dN} = \frac{-s}{2N^{1.5}}, \tag{3}$$

where s, the sample estimate of σ,

$$s = \sqrt{\frac{\sum_i (x_i - M)^2}{N - 1}} \tag{4}$$

is presumed constant with respect to N (*i.e.*, is presumed unbiased; *cf.* Dixon and Massey, 1957). We specify that a subject be terminated when the change in the dependent variable is less than the above quantity times the number of sessions to date. Since N can change only in integral steps, the critical point will occur halfway between two values of N. We therefore add 0.5 to N, to get the upper category boundary of which N is the midpoint. Our criterion becomes:

$$\Upsilon = \frac{sJ}{2(N + 0.5)^{1.5}}. \tag{5a}$$

For treatment of groups as a whole, it is:

$$\Upsilon' = \frac{sJ}{2N(N + 0.5)^{1.5}}. \tag{5b}$$

Let us see how this works for the data of Cumming and Schoenfeld. Equation 5a instantiates:

$$\Upsilon = \frac{15.7J}{2(6.5)^{1.5}} \tag{6}$$

$$\Upsilon = .47J \tag{7}$$

The criterion may be represented as a straight line that changes in error from one session to the next (dE/dJ) must cross below. The changes in error are equivalent in magnitude to changes in the dependent variable, but since it is not known *a priori* whether such changes represent an increase or decrease in error, we conservatively require that the absolute value of the change be less than the criterion. The number of sessions required for each of Cumming and Schoenfeld's subjects to pass this criterion for six days in a row ranged from 14 to 35, and averaged 23. The average rates over those days deviated from the asymptotic rates by 18%—a performance within the range established by the first two approaches (see Table 1).

Our criterion was generated in reference to group error, not individual error, and may be evaluated on that basis. The average response rate over animals during the stable conditions was 48.7 responses per minute. This deviates

by less than 6% from the average asymptotic rate of 46.1 responses per minute, and is easily included in the 95% confidence interval for the mean.

The nomothetic criterion requires an estimate of the standard deviation of asymptotic response rates, but this is not a severe problem. Let us assume a large error in estimating the standard deviation. Had we presumed it to be 7.9 instead of 15.7, we would have found ourselves conducting an average of 34 sessions, 11 more than previously. An underestimation of the sampling variability thus biases us toward running additional sessions, since the marginal utility of an extra animal becomes less than the marginal utility of an extra session, until additional sessions with their decreasing marginal returns and cumulating costs have again balanced the scales. Conversely, had we presumed the standard deviation to be 31.4 instead of 15.7, we would have run an average of 17 sessions instead of 23. The subjects' average response rate when stopped would have been 45.7—an error no greater than that occurring with the stricter criterion, although the average deviation for individual animals increases to 21%.

Problems of estimation may be further alleviated by use of the coefficient of variation, rather than the standard deviation, in the above calculations. The coefficient of variation is the standard deviation divided by the mean; it is less variable over subjects—and *a fortiori* over experiments—than is the standard deviation (*cf.* Cumming and Schoenfeld, 1960, P. 73). Cumming and Schoenfeld's coefficient of variation for average asymptotic response rates was 31%, while their within-subject coefficient of variation averaged 14.3% over the last 100 sessions. It is unlikely that the between-subject variability will ever be less than the within-subject variability, so that when the former is lacking, the latter may be taken as a conservative estimate of it. If the coefficient of variation is used in Equation 5a, the left side of the equation must also be divided by the mean, so that we test proportional change in error, rather than absolute change, against the criterion. Since the coefficient of variation (V) is traditionally based on the sample standard deviation, rather than the estimate of population standard deviation (*i.e.*, it employs a divisor of N, rather than $N - 1$), our criterion becomes

$$\Upsilon = \frac{VJ}{2(N + 0.5)^{1.5}} \left(\frac{N}{N-1}\right)^{0.5}, \tag{8}$$

or, approximately

$$\Upsilon = \frac{VJ}{2N^{1.5}}. \tag{9a}$$

When the group will be terminated as a unit, Equation 5b transforms to approximately:

$$\Upsilon' = \frac{VJ}{2N^{2.5}}. \tag{9b}$$

Discussion

It may be argued that sampling error over subjects is somehow less important than failure to reach asymptote. For instance, the former may be expected to be random, whereas the latter may be biased, with all individuals approaching asymptote from the same direction. In fact, this was not a problem in the analysis of Cumming and Schoenfeld's data, where our estimate of the terminal rate was 6% too high, even though all animals approached asymptote from below. But the argument has some merit, and is applicable to any type of stability criterion. If experimenters do not sequence conditions according to a Latin Square design, or pretrain animals to a range of response rates, they may wish to adjust the stability criterion. If sampling error is considered to be only half as detrimental as deviation of individual animals from asymptote, the experimenter merely need halve the criterion value. Similar proportional changes in the criterion will accommodate different evaluations of the relative cost of subjects and sessions. Whatever the experimenter's decision, the value of Υ chosen provides an explicit statement of the probable error in the data, which is an important advance over visual stability criteria.

To employ Equations 9a or 9b, we need merely plot a straight line on graph paper, with origin at zero and a slope of Υ or Υ'. The abscissa will measure sessions, and the ordinate will measure the proportional rate of change in error, as estimated by the differences in the dependent variable from one day to the next divided by the mean value of the dependent variable on those days. To increase the smoothness of these "operating characteristics", I graph the average of the previous day's entry and the current change index (this latter tactic generates an exponen-

tially-weighted moving average of the change indices). Thus, if D_i is the current value of the dependent variable (for each subject when using 9a or averaged over subjects when using 9b), D_{i-1} the value of the dependent variable on the previous session, and Y_i the value to be graphed, then

$$Y_i = 0.5(Y_{i-1} + | D_{i-1} - D_i | / D_i). \quad (10)$$

Inspection of Equations 9a and 9b reveals that the criterion line becomes steeper, and thus easier to pass, as the number of subjects is decreased. This seems to penalize the researcher who employs a large number of subjects, by requiring that they be run for a greater number of sessions than is required of a researcher who employs fewer subjects. This outcome follows from the logic of the nomothetic approach, where the choice of N is not arbitrary, but is presumed to be based both on the variability expected between subjects, and on the speed with which the dependent variable is expected to approach asymptote. Increases in N betoken, in a balanced allocation, increased resources, and Equations 9a and 9b automatically apportion some of those resources to an increased number of sessions. If a researcher finds the present Y criterion too lenient, it merely implies that the researcher is using too few subjects for the amount of resources that he or she is willing to invest.

It is usually necessary to shape the animals to respond, or pretrain them in other ways, before one can even begin to measure the dependent variable. These initial costs may be taken into account by allocating them to each of the conditions to be run, and then appropriately offsetting the origin of the change indices. If, for example, 12 days of pretraining are necessary for a proposed ABA design, we should begin graphing the change indices four days to the right of the origin.

At the heart of the nomothetic approach is our desire to optimize the information that we get from any experiment about population characteristics. This goal is equally valid whether we employ within-group or between-groups designs. In both cases, we take as N the number of subjects in each condition—whether or not the same subjects have appeared in other conditions. The logic of when to use each type of design has recently been reviewed by Greenwald (1976; see also Erlebacher, 1977). The within-subject design is appealing, when appropriate, because we can usually decrease the variability in our dependent variables by choosing as our dependent variable the *difference* or *ratio* of each individuals' performance in the various experimental conditions, rather than the absolute values of their performance. This decrease in V translates into decreases in the number of subjects needed, and, given the constraint of fixed resources, permits each to be run for a greater number of sessions.

GENERAL DISCUSSION

Three approaches to the evaluation of a stability have been formulated. The learning-curve approach was rejected as unwieldy, even though it performed about as well as the subsequent techniques. The trend analysis was simpler to apply, and provides researchers the flexibility of experimenting with their own cut-points and number of sessions to be tested. The nomothetic approach addressed the issue of behavioral stability at a more fundamental level. In this approach, the importance of minimizing both learning error and subject sampling error was assumed. Some such assumption is implicit in most research, whose results are inevitably generalized beyond the three or four subjects employed. In the nomothetic approach, representativeness is measured, not assumed. As number of subjects is increased, representatives of the sample is increased, with rate of increase being a function of the number of subjects already in the sample, and the intersubject variability typical of the paradigm. Just as there is a decrease in the marginal utility of each extra subject, so there is a decrease in the marginal utility of each extra session of training. When the latter exceeds the former, it is time to add new subjects—or to stop the experiment, if N is large enough. This logic formed the basis for the nomothetic approach, and led to the criterion lines of Equations 9a and 9b.

Some researchers may insist that they are interested only in the idiosyncracies of the few animals in their experiment. Such idiographic research needs no normative criteria, and may indeed be terminated at the discretion of the investigator. Other researchers may wish to discriminate between two or more theories, and the stopping criteria we proposed,

while optimal, may not be adequate. But it is short-sighted to attend to only one source of variability (deviation from asymptote) while ignoring other sources; the essence of theories is their generality over a population, and researchers interested in testing theories must be willing to invest the necessary resources, in both subjects and sessions. Conversely, for simpler questions—such as, "is there any effect from treatment B"—we may employ fewer subjects, and stop them short of asymptote.

The estimation of parameters has been a central problem of statistics for some time. The approaches evaluated in the present paper are not new. Bush (1963) estimated the number of sessions necessary to maximize the precision of estimates of the parameters for single-operator linear learning models. Although restricted to a single model (similar to Equation 1), his apportioning of resources between subjects and sessions adumbrates the nomothetic philosophy. Kazdin (in Hersen and Barlow, 1976) evaluated the statistics available for "*N* of 1" research. These statistics are essentially trend analyses, with major trends removed, so that the residuals are uncorrelated and become amenable to analyses of variance. A provocative discussion of such approaches is provided by Michael (1974). Wald (1947) and others have refined the theory of sequential statistical tests, in which sampling is continued until the ratio of positive to negative outcomes passes one of two or more boundaries. These "sequential probability ratio tests" are primarily useful for data whose parameters do not change as a function of trials, but the associated graphical technique, wherein the sampling is stopped when an operating characteristic curve passes a boundary line, is similar to the nomothetic tests outlined above. Finally, R. A. Fisher's (1935/1953) emphasis on maximizing the precision (defined as the reciprocal of the variance) of a parameter estimate for a fixed experimental cost indicates a concern for efficiency that is fundamental to the present analyses.

In Experiment I, we were concerned with a model of the learning curve, in Experiment II with a model of "visual stability tests", and in Experiment III with a model of an efficient experimenter, one committed to representativeness as well as stability. The trend test formulated in Experiment II captures the criteria that derive from current laboratory contingencies (Polya, 1954; Skinner, 1956, 1958). The nomothetic analysis questions those contingencies. It suggests that the issues of precision, cost-effectiveness, and representativeness are properly involved in every scientific decision. They deserve explicit consideration in experimental analyses of behavior.

REFERENCES

Anderson, N. H. Comparison of different populations: Resistance to extinction and transfer. *Psychological Review,* 1963, **70,** 162-179.

Bush, R. R. Estimation and evaluation. In R. D. Luce, R. R. Bush, and E. Galanter (Eds), *Handbook of mathematical psychology.* New York: Wiley, 1963. Pp. 429-469.

Culler, E. and Girden, E. The learning curve in relation to other psychometric functions. *American Journal of Psychology,* 1951, **64,** 327-349.

Cumming, W. W. and Schoenfeld, W. N. Behavior stability under extended exposure to a time-correlated reinforcement contingency. *Journal of the Experimental Analysis of Behavior,* 1960, **3,** 71-82.

Dixon, W. J. and Massey, F. J., Jr. *Introduction to statistical analysis.* New York: McGraw-Hill, 1957.

Erlebacher, A. Design and analysis of experiments contrasting the within- and between-subjects manipulation of the independent variable. *Psychological Bulletin,* 1977, **84,** 212-219.

Estes, W. K. Toward a statistical theory of learning. *Psychological Review,* 1950, **57,** 94-107.

Falk, J. L. Issues distinguishing idiographic from nomothetic approaches to personality theory. *Psychological Review,* 1956, **63,** 53-62.

Fisher, R. A. *The design of experiments.* New York: Hafner, 1935/1953.

Greenwald, A. G. Within-subjects designs: To use or not to use? *Psychological Bulletin,* 1976, **83,** 314-320.

Hersen, M. and Barlow, D. H. *Single case experimental designs: Strategies for studying behavior change.* New York: Pergammon, 1976.

Michael, J. Statistical inference for individual organism research: mixed blessing or curse? *Journal of Applied Behavior Analysis,* 1974, **7,** 647-653.

Polya, G. *Patterns of plausible inference.* Princeton: University Press, 1954.

Robertson, T. B. *Principles of biochemistry.* New York: Lea & Ferbiger, 1920.

Sidman, M. *Tactics of scientific research.* New York: Basic Books, 1960.

Skinner, B. F. A case history in scientific method. *American Psychologist,* 1956, **11,** 221-233.

Skinner, B. F. The flight from the laboratory. In J. T. Wilson *et al.* (Eds), *Current trends in psychological theory.* Pittsburgh: University Press, 1958. Reprinted in A. C. Catania's (Ed), *Contemporary research in operant behavior.* Glenview: Scott, Foresman, 1968.

Stevens, J. C. and Savin, H. B. On the form of learning curves. *Journal of the Experimental Analysis of Behavior,* 1962, **5,** 15-18.

Wald, A. *Sequential analysis.* New York: Wiley, 1947.

Received 6 August 1976.
(Final acceptance 1 July 1977.)